Supplemental Experimental Procedures

eQTL Calling and SNP Selection

RNA Mapping
Illumina 100 bp paired-end RNA-seq reads from 446 lymphoblastoid cell lines (LCLs) sequenced by the Geuvadis consortium were downloaded from http://www.ebi.ac.uk/ena/data/view/ERP001942 (Lappalainen et al., 2013). Samples were mapped with Tophat v2.0.9 without coverage searching and an edit distance of 3 against human genome build 37 using Gencode v13 as a transcriptome guide (Harrow et al., 2012; Kim et al., 2012). Coverage across genes were estimated with Cufflinks v2.1.1 using multi read correction and masking of rRNA and tRNA loci (Trapnell et al., 2010). Fragments per kilobase of transcripts (FPKM) estimates for each gene were acquired from Cufflinks and genes were filtered for having at least one individual with an FPKM of 0.5 or greater. Expression values were log2 transformed and normalized for both known and hidden covariates using the PEER software package (Stegle et al., 2010). Gender, population and the sequencing laboratory for each sample were provided to PEER as known covariates and the number of unknown covariates was set to 10.

Sample Imputation
Genotypes were obtained for 420 RNA-seq samples from the phase 1 release of the 1000 genomes (1KG) project (Consortium). For the remaining 26 samples where 1KG phase1 data was not available, we obtained Illumina OMNI 2.5M, Affymatrix Axiom 5M and HapMap data. Merged genotypes for the 26 individuals were imputed against the full phase 1 collection of genotypes using IMPUTE2 with the requirement of an imputation confidence score of 0.9 or greater for keeping the imputed call. All indels greater than 35 bp were removed from subsequent analysis (Howie et al., 2009).

eQTL Calling
Genotypes were separated into three groups by population: all (1KG), Yoruba (YRI) and European (EUR) individuals, and filtered to include only variants at greater than 5% minor allele frequency within each population group. PEER residuals and genotypes were provided to matrix eQTL to calculate cis eQTLs (SNP/gene distance less than 1Mb) using an additive linear model. To set a significance threshold we ran 1000 permutations for chromosomes 1,7,16 and 19 in each population group and calculated the p-value that corresponded to an empirical 0.1% false positive rate and used these values as significance thresholds in our eQTL analysis. This resulted in p-value cutoffs of $6.3x10^{-11}$ and $4.1x10^{-10}$ for YRI and EUR respectively (eQTL count: 471 in YRI & 3171 in EUR). For every gene with a significant snp/gene association, we performed a conditional analysis for all other SNPs within the 1Mb window of the gene using the top associated variant as a covariate in the linear model and the same p-value thresholds used in the initial pass. We iterated through consecutive rounds of conditional analysis adding each new top associated SNP to the linear model until no other variants in the region showed a significant association with gene expression (conditional eQTL count: 315 in EUR & 8 in YRI).

SNP Selection for Reporter Assay
To select variants for testing in the 79k reporter assay we first selected the top associated variant for each significant eQTL in both the primary and conditional eQTL analysis of EUR and YRI (3,965). We calculated linkage disequilibrium (LD) for every top association within the discovery population and selected all variants that were in perfect LD (r^2=1) with it (12,321 variants). In addition, for each gene significantly associated in EUR we selected the top associated variant and all variants in perfect LD within the YRI and 1KG analysis regardless of the strength of the association (12,230 variants, 4,177 redundant with prior selections).

We then selected a subset of eQTL peaks to characterize comprehensively, beyond only the top associated variants. We chose 209 eQTLs for testing of all variants having an r^2 of 0.9 or greater with the most highly associated variant (9,921 variants, 1,122 redundant with prior selections). We selected these 209 peaks based on their intersection with SNPs in the NHGRI's catalog of published GWAS hits, capturing a total of 163 GWAS SNPs. To determine overlapping peaks, we calculated r^2 within the 1KG EUR supergroup to top associated variants in the eQTL and GWAS analysis. We called loci overlapping when a shared variant was within 0.8 r^2 of the GWAS loci and 0.9 r^2 of the eQTL peak.

To create oligonucleotides (oligos) for the 29,173 sites we centered the variant within 150 bp of flanking hg19 sequence (74 bp on the 5' and 75 bp on the 3' end for SNPs). To determine the orientation of the oligo we chose the direction of the variant relative to the transcription start site of the gene it was associated with in the eQTL analysis. If the variant was associated with multiple genes in different orientations, we designed oligos in both the forward and reverse direction. For variants within the test set where other variants fell within the 150 bp oligo, we created an additional alternative haplotype oligo testing the alternate and reference allele of the centered variant with the alternative allele(s) inserted into the flanking sequence. Finally, there were 7 variants that contained an AsiI restriction site within the flanking sequence that would cause these oligos to be lost during construction of the MPRA library. To rescue the oligos we made a single base change in the restriction site; none of these changes altered the variant site or 20 bp flanking either side. In total we designed 78,956 oligos for synthesis, testing a total of 29,173 variants.

The smaller 7.5k oligo reporter library was constructed of 5 subsets of variants (2 positive controls and 3 negative controls). For positive controls we randomly selected 137 variants that were expression positive but emVar negative and 127 emVar positive variants from the 79k oligo experiment. For the location matched negative controls we compiled a list of all variants with a MAF >= 5%, residing 150-1000 bp away from a top eQTL association in the EUR analysis. Variants were filtered for having low LD with the lead association (<= 0.25 r^2) and for having no detectable signal of eQTL association in both the Geuvadis LCL dataset (p-value >= 0.001) and 13 tissues from the GTEx consortium (p-value >= 0.0001). If multiple variants met this criteria at any one loci, only a single site was selected at random for testing. For the randomly selected control variants we picked 2700 sites at random from the EUR population matching the allele frequency distribution of lead variants in the primary EUR eQTL analysis. For 1200 of the 2700 variants we set an additional criteria on requiring eQTL signal in LCLs and GTEx at the same thresholds as the location matched controls. Downstream analysis suggested not differences between the two randomly selected control sets prompting us to combine all 2700 sites together as a single set. All sites were tested in the forward orientation with the flanking sequence taken from hg19. There were 572 additional variants designed on the assay that were unrelated to the positive and control sets that were discarded from the primary analysis.

Massively Parallel Reporter Assay

Oligo Synthesis

Oligos were synthesized by Agilent Technologies as 180 bp sequences containing 150 bp of genomic context and 15 bp of adapter sequence at either end (5'ACTGGCCGCTTGACG [150 bp oligo] CACTGCGGCTCCTGC3') (Figure 1A). Post synthesis (Figure 1B), 20 bp barcodes and additional adapter sequences were added by performing 28 emulsion PCR reactions each 50 µL in volume containing 1.86 ng of oligo, 25 µL of Q5 NEBNext MasterMix (NEB, M0541S), 1 unit Q5 HotStart polymerase (NEB, M0493S), 0.5 uM MPRA_v3_F and MPRA_v3_20I_R primers and 2 ng BSA (NEB, B9000). PCR master mix was emulsified by vortexing with 220 µL Tegosoft (Evonik), 60 µL ABIL WE (Evonik) and 20 µL Mineral Oli (Sigma, M5904) per 50 µL PCR reaction at 4°C for 5 min. 100 µL of Emulsion mixture was plated per well across a 96 well plate and cycled with the following conditions; 95°C for 30 sec, 15 cycles of (95°C for 20 sec, 60°C for 10 sec, 72°C for 15 sec), 72°C for 5min. Amplified emulsion mixture was broken and purified by adding 1 mL of 2-butanol (VWR, AA43315-AK), 50 µL of AMPure XP SPRI (Beckman Coulter, A63881) and 80 µL of binding buffer (2.5M NaCl, 20% PEG-8000) per 350 µL of Emulsion mix and vigorously vortexing followed by incubation for 10 minutes at room temperature. Broken emulsion/butanol mixture was spun at 2900 rcf for 5 min and the butanol phase was discarded. The aqueous phase was placed on a magnetic rack for 20 minutes prior to aspiration. Remaining beads were washed once with 2-butanol, three times with 80% EtOH and eluted in EB (Qiagen, 19086).

MPRA Vector Assembly

To create our mpraΔorf library (Figure 1C), barcoded oligos were inserted into sfiI digested pGL4:23:ΔxbaΔluc by gibson assembly (NEB, E2611) using 1.1 µg of oligos and 1 µg of digested vector in a 40 µL reaction incubated for 60 min at 50°C followed by SPRI purification and elution in 20 µL of EB. Half of the ligated vector was then transformed into 100 µL of 10-beta e.coli (NEB, C3020K) by electroporation (2kV, 200 ohm, 25 µF). Electroporated bacteria were immediately split into eight 1 mL aliquots of SOC (NEB, B9020S) and recovered for 1 hour at 37°C then independently expanded in 20 mL of LB supplemented with 100 µg/mL of carbenicillin (EMD, 69101-3) on a floor shaker at 37°C for 6.5 hours. After outgrowth aliquots were pooled prior to plasmid purification (Qiagen, 12963). For each of the aliquots we plated serial dilutions after SOC recovery and estimated a library size of $>10^8$ CFUs.

To create our final mpra:gfp library (Figure 1D), 20 µg of mpra:Δorf plasmid was linearized with 200 units of AsiSI (NEB, R0630) and 1x cutsmart buffer (NEB) in a 500 µL volume for 3.5 hours at 37°C. An amplicon containing a minimal promoter, GFP open reading frame and a partial 3' UTR was then inserted by gibson assembly using 10 µg of AsiSI linearized mpraΔorf plasmid, 33 µg of the GFP amplicon in 400 µL of total volume for 90 minutes at 50°C followed by a 1.5x beads/sample SPRI purification. The total recovered volume was digested a second time to remove remaining uncut vectors by incubation with 50 U of AsiSI, 5 U of RecBCD (NEB, M0345), 10 µg BSA, 1 mM ATP, 1x NEB Buffer 4 in a 100 µL reaction for 6 hours at 37 C followed by SPRI purification and elution with 55 µL of EB.

To generate transfection ready MPRA libraries 10 µL of mpra:gfp plasmid was electroporated (2kV, 200 ohm, 25 µF) into 220 µL of 10-beta cells. Electroporated bacteria was split across 6 tubes and each recovered in 2 mL of SOC for 1 hour at 37°C then added to 500 mL of LB with 100 µg/mL of carbenicillin and grown for 9 hours at 37°C prior to plasmid purification (Qiagen, 12991). We repeated this same electroporation protocol 3 additional times, each time with an estimated transformation efficiency of $>10^8$ cfu. All plasmid preps were then pooled and normalized to 1 µg/uL to generate our final mpra:gfp library used in all subsequent transfections.

MPRA Transfections

Lymphoblastoid cells were grown in RPMI (Life Technologies, 61870) supplemented with 15% FBS (Life Technologies, 26140) maintaining a cell density of 2-10x10^5 cells per mL. For all 8 transfections (5 x NA12878 and 3 x NA19239) cells were grown to a

density of ~$1x10^6$ cells/mL prior to the removal of $5x10^8$ cells. Cells were collected by centrifugation at 120x g and eluted in 4 mL of RPMI with 500 µg of mpra:gfp library. Electroporation was performed in 100 µL volumes with the Neon transfection system (Life Technologies) applying 3 pulses of 1200 V for 20 ms each. Using separate control transfections we achieved transfection efficiencies of 40-60% for all replicates. Cells were allowed to recover in 180 mL in RPMI with 15% FBS for 24 hours then collected by centrifugation, washed once with PBS, collected and frozen at -80°C (Figure 1E).

Hepatocytes were grown in MEM alpha (Life Technologies, 32561) supplemented with 10% FBS. Cells were plated across ten 15 cm cell culture plates and grown to 60-70% cell density. On the day of transfection media was replaced with 30 mL fresh MEM/FBS followed by transfection with 87.5 µL of Lipofectamine 3000 (Life Technologies, L3000015) and 35 µg of DNA using the manufacturer's protocol. Cells were incubated with transfection reagents for 24 hours, then washed with 15 mL of PBS followed by dissociation with 0.05% trypsin-EDTA (Life Technologies, 25300), centrifugation, PBS wash and a final collection at 300x g prior to storage at -80°C.

RNA Extraction and cDNA Synthesis

Total RNA was extracted from cells using Qiagen Maxi RNeasy (Qiagen, 75162) following the manufacturer's protocol including the on-column DNase digestion. A second DNase treatment was performed on the purified RNA using 5 µL of Turbo DNase (Life Technologies, AM2238) in 750 µL of total volume for 1 hour at 37°C. The digestion was stopped with the addition of 7.5 µL 10% SDS and 75 µL of 0.5M EDTA followed by a 5 minute incubation at 70°C. The total reaction was then used for pulldown of GFP mRNA. Water was added to the DNase digested RNA to bring the total volume to 898 µL to which 900 µL of 20X SSC (Life Technologies, 15557-044), 1800 µL of Formamide (Life Technologies, AM9342) and 2 µL of 100 uM biotin-labeled GFP probe (GFP_BiotinCapture_1-3, Table S3) were added and incubated for 2.5 hours at 65°C. Biotin probes were captured using 400 µL of pre-washed Streptavidin beads (Life Technologies, 65001) eluted in 500 µL of 20X SSC. The hybridized RNA/probe bead mixture was agitated on a nutator at room temperature for 15 minutes. Beads were captured by magnet and washed once with 1x SSC and twice with 0.1x SSC. Elution of RNA was performed by the addition of 25 µL water and heating of the water/bead mixture for 2 minutes at 70°C followed by immediate collection of eluent on a magnet. A second elution was performed by incubating the beads with an additional 25 µL of water at 80°C. A final DNase treatment was performed in 50 µL total volume using 1 µL of Turbo DNase incubated for 60 minutes at 37°C followed by inactivation with 1 µL of 10% SDS and purification using RNA clean SPRI beads (Beckman Coulter, A63987).

First-strand cDNA was synthesized from half of the DNase-treated GFP mRNA with SuperScript III and a primer specific to the 3' UTR (MPRA_v3_Amp2Sc_R) using the manufacturer's recommended protocol, modifying the total reaction volume to 40 µL and performing the elongation step at 47°C for 80 minutes. Single-stranded cDNA was purified by SPRI and eluted in 30 µL EB.

Tag-seq Library Construction

To minimize amplification bias during the creation of cDNA tag sequencing libraries, samples were amplified by qPCR to estimate relative concentrations of GFP cDNA using 1 µL of sample in a 10 µL PCR reaction containing 5 µL Q5 NEBNext master mix, 1.7 µL Sybr green I diluted 1:10,000 (Life Technologies, S-7567) and 0.5 uM of TruSeq_Universal_Adapter and MPRA_Illumina_GFP_F primers (Table S3). Samples were amplified with the following conditions: 95°C for 20 seconds, 40 cycles (95°C for 20 sec, 65°C for 20 sec, 72°C for 30 sec), 72°C for 2 min. All LCL cDNA samples had a cycle threshold of approximately 11 while HepG2s showed an earlier cycle threshold corresponding to the larger amount of RNA recovered. To add Illumina sequencing adapters, cDNA samples and 5 mpra:gfp plasmid controls were diluted to match the replicate with the lowest concentration and 10 µL of normalized sample was amplified using the reaction conditions from the qPCR scaled to 50 ul, excluding Sybrgreen I and using only 10 amplification cycles. Amplified cDNA was SPRI purified and eluted in 40 µL of EB. Individual sequencing barcodes were added to each sample by amplifying the entire 40 µL elution in a 100 µL Q5 NEBNext reaction with 0.5 uM of TruSeq_Universal_Adapter primer and a reverse primer containing a unique 8 bp index (Illumina_Multiplex) for sample demultiplexing post-sequencing. Samples were amplified at 95°C for 20 seconds, 6 cycles (95°C for 20 sec, 64°C for 30 sec, 72°C for 30 sec), 72°C for 2 minutes. Indexed libraries were SPRI purified and pooled according to molar estimates from Agilent TapeStation quantifications. Samples were sequenced using 1x30 bp chemistry on an Illumina HiSeq through the Broad Institute's walk-up sequencing service.

To determine oligo/barcode combinations within the mpra pool, Illumina libraries were prepared from the mpraΔorf plasmid library by performing 4 separate amplifications with 200 ng of plasmid in a 100 µL Q5 NEBNext PCR reaction containing 0.5 uM of TruSeq_Universal_Adapter and MPRA_v3_TruSeq_Amp2Sa_F primers with the following conditions: 95°C for 20 sec, 6 cycles (95°C for 20 sec, 62°C for 15 sec, 72°C for 30 sec), 72°C for 2 minutes. Amplified material was SPRI purified using a 0.6x bead/sample ratio and eluted with 30 µL of EB. Sequencing indexes were then attached using 20 µL of the eluted product and the same reaction conditions as for the tag-seq except the number of enrichment cycles were lowered to 5. Samples were molar pooled and sequenced using 2x150 bp chemistry on Illumina HiSeq and NextSeq instruments through the Broad Institute's walk-up sequencing service.

7.5k Oligo MPRA Experiment
To perform the MPRA experiment of the 7.5k library we adjust the experimental conditions to 1/10th the scale used for the 79k library with the following exceptions. The two gibson assembly steps were performed at ¼ scale of the original library and RNA extraction was performed using Qiagen Midi RNeasy (Qiagen, 75142) followed by ½ scale reactions for the GFP pulldown and cDNA synthesis. Library preparation was performed as described above but diluting the samples to the cycle threshold specific for the 7.5k library.

Single Oligo Validation
To validate the expression values obtained by MPRA we selected 29 oligos consisting of 13 ref/alt pairs and 3 additional oligos. We selected the oligos based upon their association with an eQTL/GWAS peak while maintaining diverse representation of regulatory strength. Two of the oligos were selected as no-expression controls for having uncorrected p-values of greater than 0.01. We designed the same 150 bp sequence that was tested by MPRA as a gBlock (IDT) and cloned each into the pGL4.23 firefly luciferase reporter vector (Promega, E8411). We initially performed a standard luciferase reporter assay, co-transfecting $1x10^6$ LCLs (NA12878) with 1ug of the cloned firefly luciferase vector and 200 ng of a renilla luciferase control vector (pGL4.74, Promega, E6921) and recovered for 24 hour in a 96 well plate. We performed three separate experimental replicates each with two transfection replicates per experiment for a total of six replicates per oligo. Firefly luciferase luminescence for each well was normalized to the renilla luciferase luminescence for the same well and each experiment was normalized as a log-ratio value relative to the mean of the two no-expression control oligos.

Initial analysis of the luciferase expression strength to the MPRA expression values showed moderate correlation for a portion of the oligos but also displayed discordant values for many sites. Under the hypothesis that some oligos carry novel transcription start sites causing out of frame transcription of the luciferase loci, we performed qPCR as the end point for the luciferase assay instead of luminescence. Using the same transfection protocol as the luciferase assay, we performed two transfection replicates for each oligo and extracted mRNA 24 hours post transfection using the MagMAX 96 Total RNA isolation kit (Life Technologies, AM1830). DNA was digested by incubation with 2U of Turbo DNase for 60 minutes followed by RNA SPRI purification. qPCR was performed in replicate according to the manufacturer's recommendations on the purified RNA using 1-step Sybr-to-Ct and gene-specific qPCR primers (Fluc_F/Fluc_R or Rluc_F/Rluc_R) to measure both the firefly and renilla luciferase RNA levels, with pGL4.23 and pGL4.74 plasmids serving as standards for copy number calculations. For each replicate a firefly/renilla ratio was calculated as well as a log-ratio value relative to the mean of the 2 negative control samples.

Data Analysis

Barcode/Oligo Reconstruction
Paired-end 150 bp reads from the sequencing of the mpraΔorf library were merged into single amplicons using Flash v1.2.7 (flags: -r 150, -f 220, -s 10) (Magoč and Salzberg, 2011). Amplicon sequences were kept if the 5' adapter matched with a levenshtein distance of 3 or less and 2 bp at the edges of both the 5' and internal constant sequences matched perfectly. Oligo sequences from the passing reads were then mapped back to the expected oligo sequences using BWA mem version 0.7.9a (flags: -L 100 -k 8 -O 5) (Li, 2013). Alignment scores were calculated as matching bases divided by the expected oligo size and reads with alignment scores of less than 0.95 were discarded. Remaining oligo/barcode pairs were then merged and barcodes attributable to multiple oligo sequences were marked as conflicting and removed from further analysis. In total we observed 90.2 million unique barcodes in the sequencing data.

Identification of Regulatory Oligos
Reads from the tag sequencing were filtered for the inclusion of the constant sequence within the GFP 3' UTR. Specifically, a levenshtein distance of 4 or less was required within the constant sequence at the end of the tag-seq read with the two bases directly adjacent to the barcode (base 21 & 22) required to match perfectly. Barcodes were then matched with oligo sequences determined through sequencing of the mpraΔorf library and the sum of all barcodes counts within each of the 78,956 oligos were calculated.

Oligo counts from all 18 samples (5 plasmid controls, 5 NA12878, 3 NA19239 and 5 HepG2) were passed into DESeq2 and a median-of-ratios method was used to normalize samples for varying sequencing depths (Love et al., 2013). Normalized read counts of each oligo were then modeled by DESeq2 as a negative binomial distribution (NB). DESeq2 estimates variance for each NB by pooling all oligo counts across samples and fitting a trend line to model the relationship between oligo counts and observed dispersion. It then applies an empirical Bayes shrinkage by taking the observed relationship as a prior and performing a maximum a posteriori estimate of the dispersion for each oligo.The overall result is that DESeq2 can obtain an estimate for dispersion of each oligo with greatly reduced bias by pooling information from all oligos.

We then used DESeq2 to estimate the fold change estimation between the control condition (plasmid) and each of the three experimental conditions (NA12878, NA19239, and HepG2). Again, DESeq2 applies a Bayesian shrinkage on the log ratios to prevent

false positive results at the extreme ends of expression (low and high count oligos). We use Wald's test to estimate significance for expression differences between conditions and corrected for multiple hypothesis testing by Bonferroni's method accounting for 39,479 tests. We required a corrected p-value of 0.01 or less in either the reference or alternate allele in order to call a sequence as having a regulatory effect on expression.

Identification of Expression-Modulating Variants
For the identification of variants altering expression strength we considered only variants originating from sequences determined to have a regulatory effect. We calculated p-values for allelic skew by comparing the log ratios of the reference and alternate alleles using a paired t-test with independent estimation of variance and Welch's adjustment to the degrees of freedom. This test assumes normality; to evaluate normality, we calculated z-scores for the differences of the log ratios for all 39,479 ref/alt pairs and observed a distribution very similar to that expected from sampling ratios from a normal distribution (Figure S4B & C). We further validated our approach by evaluating the qq plot for variants that failed to show regulatory effects (uncorrected expression p-value > 0.01), which are not expected to have any allelic bias (Figure S4D & E). To collapse results from the two LCL experiments, we averaged both the expression ratio and allelic skew weighted by the number of replicates from each sample; p-values were combined using Fisher's method. For all samples as well as the combined LCL analysis FDR was calculated for the skew p-value using the Benjamini-Hochberg procedure.

Downsampling and Analysis of the 7.5k Oligo Library
Analysis of the 7.5k oligo MPRA experiment was performed as described for the initial 79k library applying FDR cutoffs for expression and skew calling that matched those originally used with the 79k library. Initial analysis of the full dataset showed the smaller library had greater power to detect weaker expression changes than the original library due to a 2.5-14 fold increase in the median number of barcodes tagging each oligo (Figure S3A). As expected from our previous observation of the effect of barcodes, we detected lower dispersion in the 7.5k oligo pool as estimated by DESeq2 (Figure S3B). Therefore, to better match the two libraries, we downsampled the 7.5k dataset to match the median number of barcodes representing each oligo in the 79k pool while maintaining the rank and distribution for each 7.5k experimental replicate. Specifically, we paired replicates between the two pools, calculated the ratio of the median number of barcodes per oligo in the old to its pair in the new, and sampled without replacement that proportion of the original number of barcodes for each oligo. We applied this to each replicate and repeated 500 times, each time calculating summary statistics including the number of expression positive variants and the number of emVars per subsampling. In the text, we report the mean of these 500 experiments.

This procedure caused both the dispersion and raw counts to better reflect the 79k experiment. After downsampling, we saw no loss in our ability to call both expression and emVar positive controls included in the new library compared to the full dataset with 96% of expression variants and 72% of emVars detected (96% and 71% detected in the full dataset). In addition, Correlation was strong between predicted expression effects between the two experiments. We note that despite extensive downsampling, the dispersion was still lower for the 7.5k library than in the 79k library. As a result, it is probable that the 7.5k library still maintains a slightly higher sensitivity to detect expression effects than the original 79k library resulting in conservative false positive estimates. This conclusion is supported by the lower effect size of expression positive sites detected in all three negative control sets relative to the 264 expression positive controls.

Annotations Used for Enrichment Analysis
For enrichment analysis we downloaded narrowPeak files from the ENCODE project's FTP server (Table S4). All enrichment analysis for regulatory oligos required an overlap of 1 bp or greater with an annotation at any position along the oligo. For analysis using LCL DHS regions, unless otherwise specified we took the union of all LCL regions from UW, Duke and the Unified analysis. For the annotation positive designation we required the oligo to be overlap 2 of the following 4 categories; all TF-ChIP data (including POL), LCL DHS regions, CAGE regions for NA12878 (Nucleus, Cytosol and Cell) and chromatin marks (H3k27ac,H2A.Z,H3K4me2,H3K4me3,H3K9ac). All fold enrichments are reported as odds-ratios unless otherwise stated.

For enrichments involving promoter proximity we defined transcription start sites (TSS) by analyzing the predicted transcript abundance within our mapped RNA-seq reads. Using the cufflinks generated estimation of transcript abundance we identified genes with an average FPKM of 0.5 or greater across all LCLs and selected TSSs from transcripts within 50% of the most abundant transcript's FPKM. We counted an overlap with the core promoter when the variant fell within 100 bp upstream and 50 bp downstream of the TSS. A hypergeometric test was used to evaluate significance of all emVars relative to either the proportion promoter sites in all 29k variants tested or genome wide (all variants >=5% minor allele frequency in EUR). For strongly linked GWAS associated eQTL peaks (>= 0.9 r^2) we first tested enrichment of promoter emVars relative to all other emVars falling outside these peaks. Significance was calculated by Fisher's exact test splitting all variants into two categories; those strongly linked to a GWAS SNP (r^2 of 0.9 or greater) and all non-gwas associated variants, and testing the number of emVars falling in a promoter compared to the total number of promoter variants. To verify our analysis is not confounded by the differences in how the variants were selected (variants with an r^2 of 0.9 or greater to an eQTL peak compared to perfect LD) we performed the same analysis but

using only the GWAS eQTL peaks. We split these variants into two groups based on the maximum LD value to a GWAS SNP. Using 0.9 r^2 as a cutoff we then used Fisher's exact test to calculate significance of enrichment with promoter sites.

Transcription Factor Binding Sites
We used FIMO and HOCOMOCO v9 to calculate binding scores for the reference and alternative allele in all 29k oligos (Grant et al., 2011; Kulakovskiy et al., 2012). We identified SNPs and indels where the motif had a binding p-value of $1x10^{-5}$ or less and the predicted TF showed binding in the analogous ENCODE ChIP-seq experiments. From this list we calculated the difference in binding score between the reference and alternate allele for each TF predicted to bind. Where sites had multiple predicted binding partners we selected the TF with the greatest change between the the alleles. We binned the variants based on the allele difference score and calculated the proportion these variants represented within the three classes of function (emVars, regulatory/non emVar, not significant).

Sensitivity Estimates
For the majority of eQTL peaks, only the top associated variant (and variants in perfect LD) were tested by MPRA. Therefore, an emVar may not be detected for a given eQTL peak for one of two reasons:

(1) The causal variant was not among the top associated variants for the eQTL peak and so was not tested
(2) The causal variant was tested but the MPRA assay gave a false negative

These two reasons for failure to detect an emVar in an eQTL peak correspond to two distinct sensitivity estimates, a technical sensitivity (2 alone), and the power of the study design to detect a causal variant (1 and 2 together). To estimate the power of experimental design, we first estimated the number of true positive emVars in EUR and YRI peak independently.

$$TP = (V) - N*(1-specificity)$$

Where V is the number of variants identified by the MPRA as emVars, N is the total number of variants tested across the peaks and a specificity of 99.04%. We then simulated the number of eQTL peaks explained within EUR and YRI when selecting the specified number of emVars (TP) randomly from the list of MPRA+ variants.

To estimate the MPRA's technical sensitivity, we first note that the probability of the causal variant being among those with the maximum r^2 in the peak should increase with the difference in the variance explained (Δr^2) between the top associated variant (and variants in perfect LD) and the second best association for each eQTL peak. To quantify this relationship, we fit a logistic regression to model the effect of a peak's Δr^2 and the effect size of the top associated SNP (ES) on the probability of detecting an emVar in that peak. Specifically, we fit the following regression:

$$\text{logit}(P(\text{emVar})) = \beta_2 \ln(\Delta r^2) + \beta_1 |ES| + \beta_0$$

The regression was fit using the statsmodels toolbox in Python 3.3 (https://github.com/statsmodels/statsmodels). The natural log-transformation was used to capture the expected convexity of the covariates. Using this model, we estimated the technical sensitivity as a function only of effect size alone by setting $\Delta r^2=1$, corresponding to the maximum possible separation between the top and second to top variant.

Detection of Allelic Skew in DHS and ChIP-seq Data
Detecting allelic skew in DHS and ChIP-seq datasets is challenging because mapping of short reads produces a bias toward the reference allele, confounding measurements of skew. To circumvent this problem, we constructed personal reference genomes for the maternal and paternal haplotype of each of the lymphoblastoid cell lines for which we had DHS or ChIP-seq data using the software vcf2diploid (Rozowsky et al., 2010). We then aligned DHS or ChIP-seq data to the corresponding personal reference genomes using BWA aln/samse with the default parameters (Li and Durbin, 2009). We then filtered out aligned reads with a mapping quality of 0 and obtained the set of reads overlapping heterozygous genomic variant using bedtools (Quinlan and Hall, 2010). Finally, we obtained allele counts for reads overlapping each variant by counting the reads that mapped only to the maternal or paternal reference, or that mapped with a better alignment score to one reference than the other. Alignment scores were calculated as a -1 penalty per mismatched base and a -1 penalty for each base-pair difference of an indel. Reads with an equivalent alignment score when aligned to the maternal and paternal reference were discarded.

For calculating skew in DHS, allele counts were pooled across all replicates and all LCLs that were heterozygous for a given variant. All variants with a pooled-coverage greater than 20 with a count of at least 1 for each allele were scored for DHS skew. A p-value cutoff of 0.1 was used. For figure 5C, a coverage cutoff of 10 and a p-value cutoff of 0.1 was used. In addition, variants that showed a substantial fraction of poorly mapped reads for one allele but not the other were discarded.

For calculating skew in transcription factor binding, allele counts were pooled across all LCL replicates for samples that were heterozygous for a given variant. All variants with a pooled-coverage greater than 20 for a transcription factor with a count of at least 1 for each allele were scored for TF skew. A p-value cutoff of 0.001 was used for calculating enrichment as well as for figures S5A & B.

Allelic Replacement of rs9283753

We performed CRISPR editing in LCL to alter the rs9283753 SNP at PTGER4. We used a modified version of the pX458 cas9 vector with U6:gRNA and F1 ori removed and delivered the gRNA as a 455 bp dsDNA amplicon (Mali et al., 2014; Ran et al., 2013). A guide sequence was designed that overlapped rs9283753 at positive 6 of the gRNA (Table S3). Two versions of the guide were created changing only position 6 to match the variant to be edited. Off-target cutting was assessed by *in silico* mapping with no appreciable off-target sites identified. Homology repair templates were synthesized as 150 bp PAGE purified Ultramer oligonucleotides (IDT) with a phosphorothioate bind at the predicted cutting location of cas9 (Table S3). We used two cell lines for allelic replacement; NA12878 which is homozygous for the ancestral allele (alt, T) and NA11831 which is homozygous for the derived (ref, C). We electroporated $5x10^6$ cells with 4 ug of cas9 vector, 400 ng of gRNA and 500 nM of the homology vector using the Neon system. Cells were sorted 24 hours post-transfection for GFP expression using a MoFlo Astrios EQ Cell Sorter (Beckman Coulter) at the Harvard University Bauer Core Facility.

Clonal populations of edited cells were isolated by single-cell dilution into 384-well plates. Genotypes of the clones were determined by Illumina sequencing of a the genomic regions surrounding the SNP using primers rs9283753_ILMN_F and rs9283753_ILMN_R (Table S3). Cell lysate were obtained for successful clonal populations after two weeks of growth by lysing 100 ul of cells in 100 ul of lysis buffer (100 mM KCl, 2mM EDTA, 20 mM Tris-HCl, 2% Triton-X 100, 2% TWEEN 20, 0.08 U/ul proteinase K) at 55 C for 60 minutes followed by 95 C for 10 minutes. PCR was performed on an aliquot of the lysate using Q5 Hot Start Master Mix (NEB). After amplification of the genomic region, each clonal amplicon had unique indices for sequencing added by PCR. Amplicons were sequenced on a MiSeq (Illumina) with 2x150 bp reads. Clones that showed the edited genotypes after analysis were confirmed by Sanger sequencing.

To quantify changes in expression of the PTGER4 gene after CRISPR editing, we performed qPCR comparing edited cells to wild-type cells that had undergone the same cas9/clonal expansion process. Cells were seeded at $2.5x10^5$ cells/mL 24 hours prior to RNA isolation. For RNA isolation, $7.5x10^5$ cells were collected and RNA was isolated with the MagMAX-96 Total RNA Isolation Kit (Life Technologies) according to manufacturer's instructions. cDNA for each sample was produced from 1 ug of isolated RNA using the SuperScript III First Strand Synthesis System (Life Technologies). qPCR were performed with PowerUp SYBR Green Master Mix (Life Technologies), 10 ng of cDNA, and forward and reverse primers at 500 nM each in a total volume of 10 ul. Technical triplicates were performed for each reaction. For each edited cell line two biological (independent seedings) were performed, for the negative control samples (wild-type) 4 and 3 independent clonal populations were tested for NA12878 and NA11831 respectively. Expression values for two seperate primer pairs for PTGER4 were averaged together and normalized using the ΔΔCt method using PPIA and TBP as references (Primers listed in Table S3).

References

Consortium, T. An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Grant, C., Bailey, T., and Noble, W. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics (Oxford, England) *27*, 1017–1018.

Harrow, J., Frankish, A., Gonzalez, J., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research *22*, 1760–1774.

Howie, B., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics *5*, e1000529.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. (2012). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology *14*, R36.

Kulakovskiy, I., Medvedeva, Y., Schaefer, U., Kasianov, A., Vorontsov, I., Bajic, V., and Makeev, V. (2012). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Research *41*, D195–202.

Lappalainen, T., Sammeth, M., Friedländer, M., Hoen, P., Monlong, J., Rivas, M., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Li (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *25*, 1754–1760.

Love, M., Huber, W., and Anders, S. (2013). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Magoč, T., and Salzberg, S. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics

(Oxford, England) *27*, 2957–2963.
Quinlan, A., and Hall, I. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841–842.
Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2010). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Molecular Systems Biology *7*, 522.
Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Computational Biology *6*, e1000770.
Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology *28*, 511–515.